DATA MINING APPROACH FOR DIRECT MARKETING OF BANKING PRODUCTS WITH PROFIT/COST ANALYSIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

OZAN KORKMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2017

Approval of the thesis:

DATA MINING APPROACH FOR DIRECT MARKETING OF BANKING PRODUCTS WITH PROFIT/COST ANALYSIS

submitted by **OZAN KORKMAZ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** _____

Prof. Dr. İsmail Hakkı Toroslu
Supervisor, **Computer Engineering Department** _____

Assoc. Prof. Dr. Pınar Karagöz
Co-supervisor, **Computer Engineering Department** _____

Examining Committee Members:

Prof. Dr. Ahmet Coşar
Computer Engineering Department, METU _____

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU _____

Assist. Prof. Dr. Şeyda Ertekin
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Osman Abul
Computer Engineering Department, TOBB ETU _____

Date: Sept 08, 2017

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: OZAN KORKMAZ

Signature :

ABSTRACT

DATA MINING APPROACH FOR DIRECT MARKETING OF BANKING PRODUCTS WITH PROFIT/COST ANALYSIS

Korkmaz, Ozan
M.S., Department of Computer Engineering
Supervisor : Prof. Dr. İsmail Hakkı Toroslu
Co-Supervisor : Assoc. Prof. Dr. Pınar Karagöz

September 2017, 54 pages

Nowadays, direct marketing is widely used advertisement method by many business areas such as banks. The main purposes of direct marketing are to maximize return on investment, minimize cost of promotions and reach to peak number of customers that prefer the offerred campaign. Therefore, it is necessary to collect and process huge amount of customer related data to decide questions of which customer will be offered a product, which product will be suitable to him/her and via which channel the promotion will be presented. However, because positive customer response rates are much less than negative ones, negative data instances dominate positive ones and cause imbalance in dataset. This problem makes it difficult to make a successful selection of product and channel for a promotion and therefore, brings about decrease on true predictions and total profit value while false predictions and total cost value increase. In this thesis, methods are proposed which improve profit/cost ratio to increase return on investment while increasing accuracy rate. Experiments with proposed methods

applied on a real bank dataset show very promising profit/cost ratios and accuracy rates on predicting customers with proper products and channels. Results of experiments indicate that proposed methods yield some amount of decrease on total profit value; however, since the decrease rate of total cost value is much greater than total profit one, profit/cost ratio increases.

Keywords: Bank Marketing, Direct Marketing, Data Mining, Customer Relationship Management, Classification, Clustering

ÖZ

VERİ MADENCİLİĞİ İLE BANKA ÜRÜNLERİNİN DOĞRUDAN PAZARLAMASININ KAR/MALİYET TAHLİLİ

Korkmaz, Ozan
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. İsmail Hakkı Toroslu
Ortak Tez Yöneticisi : Doç. Dr. Pınar Karagöz

Eylül 2017 , 54 sayfa

Günümüzde, doğrudan pazarlama bankalar gibi birçok ticaret alanı tarafından yaygın olarak kullanılan reklam metodudur. Ana amacı, yatırım getirisini en üst düzeye çıkarmak, promosyonların maliyetini en aza indirmek ve teklif edilen kampanyayı tercih eden en yüksek müşteri sayısına ulaşmaktır. Bu nedenle, hangi müşteriye bir ürün önerileceği, hangi ürün kendisine uygun olacağı ve promosyonun kendisine hangi kanalla sunulacağı ile ilgili karar vermek için büyük miktarda müşteri verisini toplamak ve işlemek gerekir. Bununla birlikte, olumlu müşteri yanıt oranları olumsuzlardan daha düşük olduğu için olumsuz veri örnekleri olumlu örneklere üstün gelir ve veri setinde dengesizliğe neden olur. Bu sorun, tanıtım için ürünün ve kanalın başarılı bir şekilde seçilmesini zorlaştırır. Bu nedenle yanlış öngörüler ve toplam maliyet değeri artarken gerçek tahminler ve toplam kar değerlerinde düşüş getirir. Bu tezde doğruluk oranını arttırırken, yatırım getirisini artırmak için kâr/maliyet oranını yükseltici yöntemler önerilmektedir. Önerilen yöntemler gerçek bir banka veri kümesi üzerinde uygulanmış

olup, uygun ürünleri ve kanalları kullanarak müşterileri tahmin etmede başarılı kar/maliyet ve doğruluk oranları göstermektedir. Deney sonuçları, önerilen yöntemlerin toplam kar değerinde bir miktar azalmaya sebep olduğunu göstermektedir. Fakat, toplam maliyet değeri düşüş oranı toplam kar değeri düşüş oranından çok daha yüksek olduğu için kar/maliyet oranı artmaktadır.

Anahtar Kelimeler: Banka Pazarlama, Doğrudan Pazarlama, Veri Madenciliği, Müşteri İlişki Yönetimi, Sınıflandırma, Kümeleme

To my beloved family.

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. İsmail Hakkı Toroslu and my co-supervisor Assoc. Prof. Dr. Pınar Karagöz for their guidance on this thesis work. I appreciate their full support and patience during my research. I enjoyed discussing the ideas about the research and learned so much from their motivational and educational talks.

I would like to thank my dear wife Merve Mıtık for supporting me during the work on this thesis and never leaving me alone through these years.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

CRM	Customer Relationship Management
PCL	Product, Channel and Label
PC	Product and Channel
PCM	Pulse Code Modulation
REPTree	Reduced Error Pruning Tree
API	Application Programming Interface
ROI	Return on Investment
LCV	Lifetime Customer Value
TAN	Tree Augmented Naive Bayes
MLPNN	Multilayer Perception Neural Network
SVM	Support Vector Machine
k-NN	k-Nearest Neighbor
ARFF	Attribute-Relation File Format

CHAPTER 1

INTRODUCTION

1.1 Motivation & Problem Definition

People always buy and sell something to maintain the flow of their life. To increase sales amount, a seller markets his/her product to customers and a customer tries to find the most suitable product to himself/herself. In this economical cycle, all of the people want to gain profit over these transactions.

In modern economies, concept of company is arisen and companies have become more profit oriented, which leads them to be more innovative. Therefore, marketing becomes an important research area. Although, traditional advertisement methods like mass marketing continues to be used by marketing agencies, there is no innovation and development about mass marketing methods. In addition, impression of mass marketing on customers decreases because of enormous increase on company count and their products [19]. One research shows that one American is, everyday, exposed to 3000 advertisements in television, radio, billboards, *et al.* and nearly 80% of Americans feel disturbed because of this massive advertisements [39]. Therefore, new methods are investigated to boost marketing industry, and usage of direct marketing concept rises. It is reported in 2000 that approximately 42% of companies switching marketing techniques from mass marketing to direct marketing [38].

Because of competition and focusing on profit lead companies to use systems which automatically improve business relationships with customers and increase ROI. Therefore, CRM system have been deployed in marketing industry, which

is a process of collecting and analyzing customer related data (e.g. purchase logs, transaction frequency) to increase transaction between customer and company. Using data, companies construct their marketing strategies, which increases customer loyalty, decreases switching costs and makes company one step ahead about the competition on gaining customers [15]. However, in CRM, the main aim is not only to sell a product, but also to create a customer life time value after all analyses. At this stage, direct marketing is a good way to enhance customer relationship and ROI of company.

Direct marketing is one of marketing strategy used in CRM systems. It is based on prospecting to find a way to reach customer with an advertisement of a product. There are a lot of sectors using direct marketing which are banking, finance, telecommunication, supermarket *et al.* In [6], Cohen describes the motto of bank and finance sector in direct marketing saying that "The right product to the right customer at the right time." However, it is not an easy process to promote multiple products to right customer via multiple channels. Because many business constraints with multiple set of products and channels increase the complexity of prediction. Besides, a direct marketing campaign is expected to maximize ROI which requires profit/cost analysis over dataset and therefore, more effort is needed.

Collected customer information in databases contains huge amount of customer entries, profiles, transactions and previously communicated campaign results. This increases complexity of analyzing datasets and it is not feasible to build a pattern representing an useful information for bank direct marketing. To decrease complexity of analysis bank direct marketing dataset, automated tools based on data mining techniques are used.

In this thesis, I focus on data mining techniques to analyze profit/cost constraint of bank direct marketing dataset. Using classification and clustering methods of data mining, large amount of customer data is processed, then data instances are grouped with similar characteristics and meaningful patterns are extracted to create models. Using models, prediction is applied over test dataset with not only examining accuracy rate but also profit/cost ratio defined in this thesis.

1.2 Contributions

Data mining applied on direct marketing problems has been researched for several decades [2] [7] [21] [23] [25] [26] [27] [28] [30]. On the other hand, there are some papers mentioning profit constraint in their contents [8] [29]. However, some of them does not concern with data mining while analyzing direct marketing problems with profit and cost constraints [5] [40]. Although, several researches have been done over direct marketing using data mining, only a few study has been done over profit and cost analysis on direct marketing problems using data mining [6] [16] [19] [22]. In this study, I worked on a direct marketing dataset with analyzing profit/cost ratio using data mining techniques. Therefore;

- previously used methods are implemented and profit/cost ratio analysis is applied on them.
- new methods are proposed to increase profit/cost ratio.
- while analyzing profit/cost ratio, accuracy rates are also discoursed on.

1.3 Organization of the Thesis

This thesis is organized as follows:

Chapter 2- Literature Survey summarizes previously studied works on direct marketing and its applications using data mining methods.

Chapter 3- Background expresses the details of methods and techniques that are used in this thesis.

Chapter 4- Proposed Methods for Profit/Cost Analysis on Direct Marketing of Banking Products is the section that gives the details of dataset, dataset generation, estimation of profit-cost values and proposed methods.

Chapter 5- Results and Discussions mentions about the results of experiments and their discussions.

Chapter 6- Conclusion and Feature Works concludes this thesis and explains future work ideas that will improve the studies mentioned in this thesis.

CHAPTER 2

LITERATURE SURVEY

The definition of marketing is that it is a management process with purpose of providing services to customers who are supposed to buy them. Marketing is a process that companies use to commercialize their products to customers. There are two types of marketing widely used by companies to present their products, which are mass marketing and direct marketing [25]. Almost all of companies selling products or services uses at least one of these advertisement methods.

Mass marketing generally depends on broadcasting channels which are radio, television, newspaper. They use these methods to market their products and services to the largest number of people as possible. However, mass marketing is not as effective as it is used to be. Because there are a lot of competitors in industry and number of companies with their products increase. Additionally, although the cost of mass marketing advertisement is high, the ROI values are generally not as they are expected [19].

Direct marketing is a method that customers are examined and categorized before reaching them. Each candidate customer is contacted directly and informed about product and services personally. Each product is selected by analyzing customer characteristics. The best communication method is determined according to previous analysis and customer is contacted via phone call, email, SMS and similar methods. Marketing systems using mass marketing techniques does not prefer them as it used to be and switches to direct marketing techniques increasingly [18] [27].

2.1 More About Direct Marketing

There are several definitions stated about direct marketing for a long years and there is no exact definition for it yet. One states that direct marketing is a marketing methodology which forces companies to collect and store customer information in databases and process them in an efficient techniques to decide a customer is valuable to be contacted or not for a specific product. One of the definitions focuses on information process, one of them puts emphasis on long term relationship between company and customer [3] [36]. Various proposed definitions about direct marketing and their explanations are discussed and new definition is proposed in [24].

Although, there are various definitions on direct marketing, there are three main characteristics of direct marketing [40], which are as below:

- Direct marketing is not a marketing instrument; but it is a marketing strategy.
- Direct marketing is an interactive where communication between customer and company is bilateral. A product can be proposed to customer by a company; in the meantime, a customer can contact with customer using call centers or Internet.
- Results of direct marketing methods are easier to measure than mass marketing ones. Cost of each campaign can be calculated immediately after campaigns accurately. With this information, efficiency of further campaigns can be estimated easily.

Additionally, direct marketing systems are growing and developing systems after each campaigns, because each contact which is successful or not, is saved in databases and used for further campaigns. Any characteristics about customers like age, job, education status are immediately stored and prepared to analysis. After that, various techniques are applied on datasets, which are collection of customer data, to find familiarity of buyer behavior. If it is predicted that customer yields to buy the product, then the focal point of campaign, means the product, is proposed the customer over predicted channel.

2.2 Data Mining for Direct Marketing

In direct marketing systems, the problem is that there are a lot of customer profiles, shopping transactions, previous campaign results and newly collected data of people who are candidate customers. Therefore, there is huge amount of data waiting to be processed. The solution is to use automated systems processing and analyzing this data. Data mining techniques provides solutions for this purposes. Clustering and classification techniques are used to group customers and generates patterns over models using customer characteristics [22].

In [19], generalized steps of data mining applied on direct marketing systems are described. Assume that a company sold a product P to $S\%$ people, and $(100 - S)\%$ of people did not buy P. All buyer customers in $S\%$ are retrieved from database and composes single dataset which is used in data mining. Then, it is stated that preprocessing on dataset is applied and dataset is split into two parts, which are test and training datasets. Learning algorithms are applied on training dataset and using resultant patterns with test dataset, models are evaluated. If results are satisfactory, then apply this patterns on $(100 - S)\%$ people and promote P to likely buyers. Figure 2.1 demonstrates each execution step of data mining on direct marketing system, which is inspired from [2].

The problem behind selecting dataset from $S\%$ is that all customers in this data set has positive class label and this is an imbalance dataset. Therefore, *Ling and Li* applies adding new negative entries to dataset accumulating with positive results. Because positive results are rare and precious than negative ones, they are preserved but negative ones are down-sampled from $(100 - S)\%$ [19]. In this thesis, same problem is occurred in classification implementations, where the number of negative class label is 27 times greater than positive ones. This causes imbalance in dataset and classification model always predicts non-buyers. Therefore, down-sampling is used on dataset before applying classification.

On the other hand, up-sampling is an other choice to diminish imbalance condition of dataset. However, usage of up-sampling does generally not scales pre-

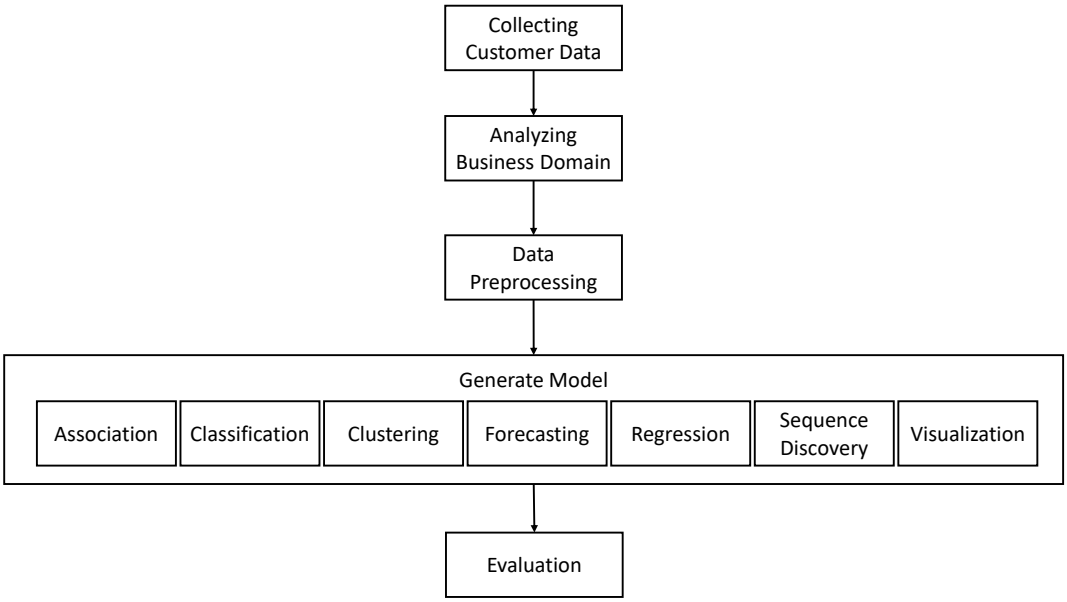

Figure 2.1: Flow of direct marketing using data mining

diction results up [17]. The reason is explained that up-sampling does not add any information to dataset but destroys singularity of it [19].

2.3 Applications of Data Mining for Direct Marketing

According to [26], 62.1% of all written articles between 2000 and 2006 about CRM are related to customer retention and 51.9% of customer retention ones are written about direct marketing. Approximately, 1 of 3 research papers on direct marketing prove that direct marketing is an important field for researchers.

There are seven types of data modeling that can be applied in data mining techniques, which are association, classification, clustering, forecasting, regression, sequence discovery and visualization [26]. However, it is important to mention that one should be careful while selecting data mining algorithms. It is important to choose the most appropriate algorithm according to dataset and business constraints [12].

In [19], similar related two problems are specified. Firstly, because of a large amount of customer and imbalanced data distribution in dataset, it is obligatory to select efficient learning algorithm. Therefore, *Ling and Li* apply ranking on

dataset and create new training dataset which keeps buyer customers data as they are; however, adds particular number of non-buyer customer data. Secondly, although acceptable patterns are generated from learning algorithms, results of prediction may not meet the requests. Because, classification errors should be handled separately with respect to business constraints. For example, false positive prediction will yield disturbance on non-buyer customer or false negative prediction will miss precious buyer customer. Additionally, splitting prediction into two judgment (buyer or non-buyer) does not provide flexibility in prediction phase. Therefore, *Ling and Li* used learning algorithms with probability measurement or certainty factor and selected Naive Bayes algorithm in their experiments.

Kohavi mentions that if dataset has large enough, then single classification algorithm does not scale accuracy up. Therefore, a new hybrid approach, called NBTree, is proposed using Naive Bayes and regular decision tree combination [17].

One of marketing datasets belongs to Portuguese bank is analyzed using some data mining algorithms by several researchers [2] [7] [23]. MLPNN, TAN known as Bayesian networks, logistic regression, and C5.0 decision tree [35] are applied on bank dataset and results shows that statistical measure results change according to learning algorithm. While MLPNN has the best result in accuracy, logistic regression gives highest result in sensitivity. On the other hand, C5.0 decision tree has the best specificity rate [7]. A different experiment on same dataset shows that MLPNN has better accuracy ratio than Naive Bayes; however, sensitivity rate of Naive Bayes learner is better than MLPNN one's [2]. Both [2] and [7] show that there is no best algorithm for all measurements. One should select the learner algorithm evaluating business constraints.

Another approach is to apply ensemble methods as a learner. Bagging is one of these methods based on applying multiple classifier on generated subset of training dataset and prediction using multiple models [4]. Other ensemble method is adaptive boosting, iterates over same classifier and after each iteration, weight of misclassified instance is increased ensuring that instance will be most prob-

ably classified correctly [11]. Ensemble methods can provide better results on prediction, therefore they are preferable to work on [2].

2.4 Profit and Cost Analysis on Direct Marketing

The fact of producing a product nearly half of its cost attracts American companies to make their products produced in outer countries [9]. All of the reason behind this outsourcing is to make more profit. Similar to outsourcing which is to get more profit by decreasing cost of production, companies prefer to promote their products via direct marketing rather than mass marketing to decrease cost of promotion. *Thomas* states that belief on having more customer and more display on market is wrong to make the product more profitable. On the contrary, marketing aiming to small subgroups of customers with planning is more profitable. Because, reaching large groups costs more and small groups needs can be analyzed separately which increases profit [39].

Knowing the response of a customer for a promotion is not enough to proceed for direct marketing. Campaign should ensure that ROI rates will be achieved and predetermined profit should be made for each product after promotions. Therefore, a framework over direct marketing is proposed including important business constraints like ROI, multiple product and channels [6]. It is mentioned that data mining applied in direct marketing generally neglects business constraints and if they are included in modeling of predictor, it will improve model performance [30]. On the contrary, campaign can be promoted to customers but there should be limit for this activity because it is shown that more investment after a certain threshold will firstly affect ROI by decreasing it, then cost of investment starts to suppress profit value [6].

There are different methods applied to increase profitability of campaign like analyzing LCV. With customer based profit analysis, the most profitable customers can be focused to increase profit; or customers with low added value to company and hurdling predefined ROI can be removed from campaign list. Companies derives their profits from 20% of their customers which is small number; there-

fore, applying customer based profitability analysis on customers is preferable [41].

Data mining algorithms used in direct marketing can be applied for profit and cost analysis. Logistic regression, MLPNN, k-NN and SVM algorithms are experimented on direct marketing dataset and results are compared with respect to profit and cost values [16].

CHAPTER 3

BACKGROUND

In direct marketing analysis using data mining, classification and clustering methods are widely used to analyze datasets. Literature review shows that neural networks, decision trees and association rules are mostly used classification methods [26]. Although, K-Means is proposed more than 50 years, it is one of the simplest and most used clustering algorithm [14].

In this thesis, classification and clustering algorithms are used over bank direct marketing dataset. To create clusters over dataset, one of improved version of K-Means algorithm, which is called K-Means++, is applied. To classify dataset instances and generate models, C4.5 Decision Tree and REPTree algorithms are used.

3.1 K-Means++

K-Means is one of clustering methods widely used. It is originally developed for vector quantization in signal processing. Main purpose of K-Means is to enhance PCM by locating quantum values more closely in voltage regions [20].

However, it is improved with some optimization steps while selecting initial k clusters. In standard K-Means, this process is based on selecting centroids arbitrarily. However, K-Means++ changes selecting initial centroid process using probability calculation. First centroid is selected randomly; however, other k-1 centroids are determined with highest probability to closest center. For $D(x)$, which is shortest distance to closest centroid, find $x \in dataset$ which has highest

$D(x)^2$ to select new centroid [1].

This study uses K-Means++ algorithm to cluster dataset into sufficient partitions.

Euclidean Distance Function

There are a lot of distance functions that can be applied on finding closest cluster centroid while predicting a customer decision, which mostly used are Euclidean and Manhattan Distance. This study uses Euclidean Distance to calculate distance between customer instance and cluster centroids.

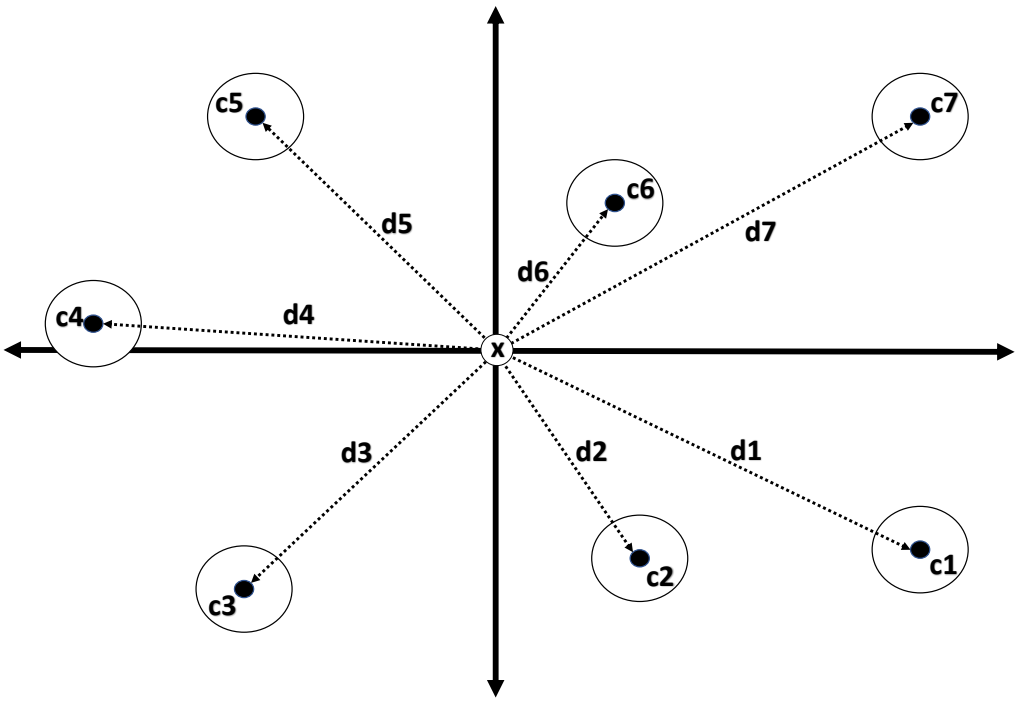

Figure 3.1: Sample data distribution after clustering

In Figure 3.1, x is a customer, $c1$ to $c7$ are cluster centroids generated by K-Means++ algorithm with $k = 7$. $d1$ to $d7$ are Euclidean distance between x and each c. The cluster with shortest distance is correlated with x.

3.2 Reduced Error Pruning Tree

REPTree is an algorithm both generates decision and regression trees by minimizing information gain and variance. It is developed to simplify decision trees

empowering it with pruning based on reduced error pruning [32]. It applies sorting operation on numeric attributes at initial stage of algorithm and no more sorting is considered rather than C4.5 Decision Tree, which sorts at each node of tree [10] [42] . Because it is a supervised learning technique, it is required a test dataset. Instead of generating multiple model trees T and selecting the best resulted one $T_{selected}$ after testing each T_i, only one tree T_{one} is constructed directly. Each subtree S, which has not leaf node, in T_{one} is tested by replacing the most proper leaf node and S. Then, T_{one} is tested again. If intermediate version of T^i_{one} after replacing has less or equal error than previous T^p_{one}, then subtree S is replaced with leaf node. This procedure continues to apply until minimal error count is achieved.

In Figure 3.2, a sample tree of regression tree model is represented, which is a small part of it. Because generated tree is big enough that it does not fit into page. Pruning operation is applied on model and the depth of tree is limited to 16. It reduces resultant regression tree complexity and overfitting problems while increasing accuracy of prediction.

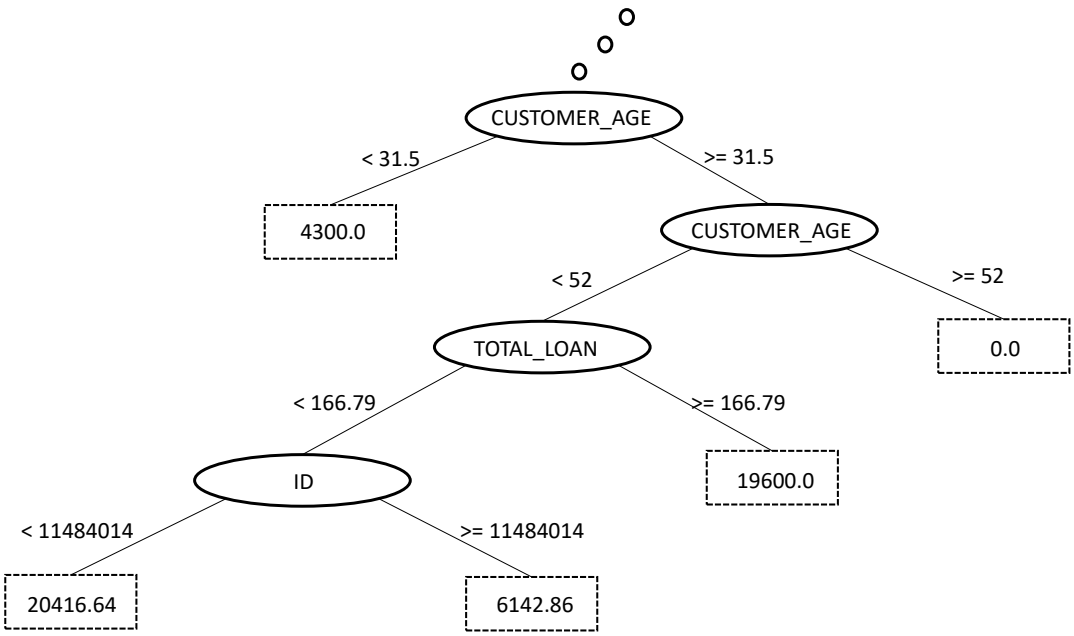

Figure 3.2: Sample regression tree using REPTree on Turkish bank dataset

3.3 C4.5 Decision Tree

Decision tree algorithms are used to divide datasets into subsets which contain instances with similar features. C4.5 Decision Tree is one of them, which is firstly proposed by Quinlan in 1993 [33]. It is denoted as the most used algorithm in data mining field [43] and successor of ID3 which is also proposed by Quinlan [31]. Processes that ID3 can not apply but C4.5 can do is that C4.5 applies pruning on subtrees and have ability to handle continuous attributes by applying binary split on them. Using divide and conquer algorithm with top-down approach on dataset D, root set D is divided into leaf subsets repeatedly using gain ratio. It finalizes when stopping condition is satisfied, which is to reach a number of instance below the threshold or zero [34]. While the tree grows, to protect the model from over-fitting, error based pruning is used to prune the tree. This pruning technique is an evaluation of pessimistic pruning and estimates error rates using confidence interval for proportions [37].

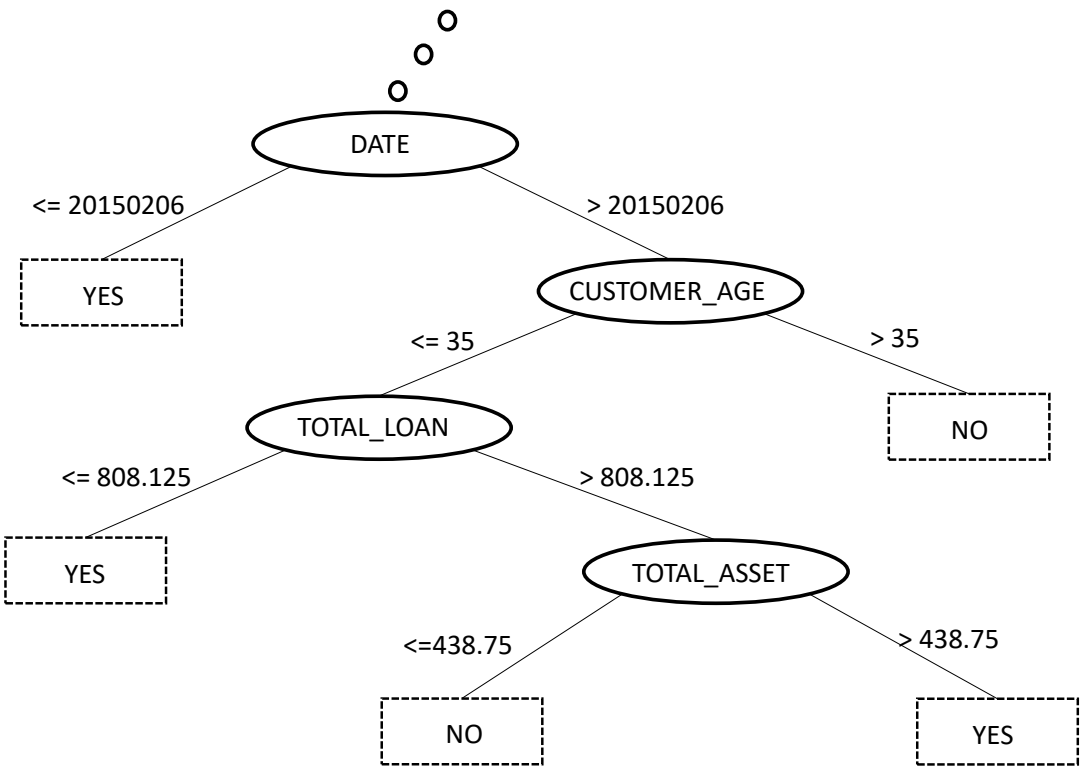

Figure 3.3: Sample decision tree using C4.5 on Turkish bank dataset

In Figure 3.3, a subtree of the C4.5 Decision tree model generated in this thesis

is represented. In final tree, leaf nodes represent class label and internal nodes are joints of pattern that lead to leaf nodes. All attributes standing for a node should be categorical; however, dataset may contain numerical values and they will be discretized at preprocessing [34].

CHAPTER 4

PROPOSED METHODS FOR PROFIT/COST ANALYSIS ON DIRECT MARKETING OF BANKING PRODUCTS

This thesis is a collaborated study worked on a bank campaign dataset, which contains two parts. The first part is based on maximizing acceptance ratio of promotions while decreasing the number of unnecessary communication, causes disturbance on customers. The second part is based on increasing profit of bank for promotions while decreasing cost of campaigns. In this thesis, second part is studied by analyzing campaign profit/cost ratio results of direct marketing dataset. In the first phase, dataset is shaped to required forms. Firstly, dataset is converted from text format to ARFF format because processing ARFF formated dataset is more easier than text one. Secondly, some data preprocessing is applied on dataset before analysis because noise that affects results negatively is not acceptable. In second phase, profit/cost ratio is tried to be increased while accuracy and sensitivity rates increase. Therefore, clustering and classification algorithms of data mining are used over dataset to find whether customer accepts promotion or not. With respect to acceptance count, profit/cost ratio and related information about it are calculated. Moreover, heuristic approaches are applied over classified model found previously.

Methods applied in this thesis use two datasets, one of them is dataset that is created after a real Turkish bank campaign and the other one is the dataset that is generated over first dataset to apply regression analysis in this thesis works.

4.1 Data Preparation

This work is partially supported by one of biggest Turkish bank within the scope of research collaboration project. Support of bank covers providing dataset of anonymous customer information.

4.1.1 Dataset

In this project, a dataset provided by one of banks in Turkey is used, which is generated using results of direct marketing campaigns. Content is based on 4 campaigns that are tried to sell to customers using 4 communication channels. Attributes of dataset related with campaigns are *Channel*, *Product* and the class *IsSold* fields. On the other side, each record represents a customer. Each customer has a unique id with some profile information. Therefore, whole dataset is built up using customer specific information and offered product and channel; and the result of the offer, whether customer accepts the promotion or not.

The dataset consists of 81915 instances with 13 fields. There are 2975 positive (yes) and 78940 negative (no) entries. The number of positive results is 3% of the total instance number of dataset, which is common drawback for a direct marketing dataset. Because positive results are important to draw a pattern when it is required to find positive rules. There exists 4 products in the dataset and the number of positive and negative responses for each product are shown in Table 4.1. These 4 products are promoted using 4 communication channels which are *SMS*, *EMAIL*, *CC* and *IVN*. Number of positive and negative responses for each channel are shown in Table 4.2.

4.1.2 Dataset Structure

Structure of the dataset used in this thesis is explained in two different table, which are Table 4.3 and Table 4.4. *Attribute Name* represents each column in the dataset, whose count is totally 13. *Attribute Description* is the explanation of each column and *Domain Values* details the range of each attribute value

Table 4.1: Details of Banking Campaign with Product Instance Counts

	Loan	Deposit Account	Credit Card	Overdraft Account
Total Count	33957	14643	14183	19132
Positive Resulted	728	749	1459	39
Negative Resulted	33229	13894	12724	19093

Table 4.2: Details of Banking Campaign with Channel Instance Counts

	SMS	EMAIL	CC	IVN
Total Count	24397	31666	15565	10287
Positive Resulted	627	1006	1272	70
Negative Resulted	23770	30660	14293	10217

which they can take.

There is an additional column that is not represented in Table 4.3 and Table 4.4, which is *Attribute Type*. The type of attributes in this dataset can be nominal or numeric.

- Nominal attributes are CHANNEL, PRODUCT, ACTIVE_CUSTOMER, EDUCATION and ISSOLD.
- Numeric attributes are ID, DATE, PERIOD, CUSTOMER_AGE, CUSTOMER_PERIOD, ACTIVE_PRODUCT, TOTAL_ASSET and TOTAL_LOAN.

4.1.3 Data Preprocessing

The Turkish Bank dataset totally holds 81939 entries, where 24 of them are duplicate. Therefore, the dataset was filtered to remove duplicates and totally

Table 4.3: Attributes of Turkish Bank Dataset, Part 1

ID	Attribute Name	Attribute Description	Domain Values
1	ID	Id of customer	[6477, 60051815]
2	CHANNEL	Which channel is used to reach customer?	{IVN, CC, EMAIL, SMS}
3	PRODUCT	Which product that contact is made for?	{Credit Card, Overdraft Account, Loan, Deposit Account}
4	DATE	Which date that contact is made?	[2015.01.13, 2015.05.31]
5	PERIOD	Which period that contact is made?	[120, 124]
6	CUSTOMER_AGE	How old is customer?	[16, 90]
7	CUSTOMER_PERIOD	How many weeks that bank works with customer?	[1, 1384]
8	ACTIVE_CUSTOMER	Is customer an active one?	{0, 1}
9	ACTIVE_PRODUCT	How many active product does customer have?	[0, 17]

used instance count decreased to 81915. Dataset included several missing values in three attributes namely *Education*, *Total Asset* and *Total Loan*. These values have been replaced with *Null* values.

Distance calculation is important to find positions of each record to cluster sets. Therefore, to decrease unnecessary calculations, I eliminated some attributes with minimum contributions using attribute selection with respect to information gain. *Customer Period*, *Id* and *Date* are selected after attribute selection

Table 4.4: Attributes of Turkish Bank Dataset, Part 2

ID	Attribute Name	Attribute Description	Domain Values
10	EDUCATION	Education status of customer	{Bachelor's Degree, Academy, Secondary School, Primary School, Master's Degree, High School, Uneducated, Doctoral Degree, Null}
11	TOTAL_ASSET	Total asset of customer	[0, 569667.025]
12	TOTAL_LOAN	Total loan of customer	[0, 450743.725]
13	ISSOLD	Has customer subscribed product?	{Yes, No}

and used for distance calculations. However, the results of attribute selection has not been logical and applying it had no significant effect. Therefore, I did not only used attributes that have maximum contribution but also the other ones.

4.1.4 Data Generation

This section explains the details of generating new dataset with label RATIO. It is used to apply regression analysis and is an enhanced version of previously dataset. Formula 4.1 explains the basic calculation of profit and cost in which P is profit function, I is income function and C is cost function where $x \in ProductList$ and $y \in ChannelList$. Before applying Formula 4.1, income of each product and cost of each channel is estimated, which is explained in Section

4.1.5 more detailed.

$$P(x) = I(x) - C(y) \tag{4.1}$$

To calculate ratio, first each customer information is parsed. Then, using his/her product and channel, profit and cost values are found by using Formula 4.1. Then, formula 4.2 is applied to change label attribute from nominal *IsSold* to numeric *Ratio*.

$$RATIO(c) = \frac{P(c)}{C(c)} \tag{4.2}$$

The main aim of changing label is to increase diversity. Product income and channel cost are not included in previous dataset. Therefore, profit/cost ratio is an enhancement over label attribute. Instead of using classes *Yes* and *No*, numerical values returned from 4.2 is used. After changing label attribute to numerical, regression tree algorithm, which is suitable for continuous values, is applied on new dataset with *Ratio* label.

Table 4.5 shows the replacement of *IsSold* attribute with enhanced *Ratio* one.

Table 4.5: New Attribute of Dataset After Enhancement

ID	Attribute Name	Attribute Description	Domain Values
13	RATIO	Profit/Cost ratio of promotion	[0.0, 42500.0]

4.1.5 Profit and Cost Estimation

To analyze profit/cost ratio, it is necessary to know the values of income for each product and cost of each channel. Basically, a bank marketing product has a constant profit when it is accepted by a customer. However, there is always a cost for each marketing move and it is communication cost in our campaign

Table 4.6: Profit of Each Product (in Turkish Liras)

	Credit Card	Overdraft Account	Loan	Deposit Account
Profit	24.5	37	42.5	10

Table 4.7: Cost of Each Channel (in Turkish Liras)

	SMS	EMAIL	CC	IVN
Cost	0.014	0.001	0.9	0.179

dataset caused by contacted channel. In this project, results of experiments are based on accuracy and profit/cost ratio. It is important to measure accuracy which provides information of correctly predicted instance ratio over incorrectly predicted ones. However, there is an important constraint that should be measured, which is profit/cost ratio. When this ratio increases, income of promotion increases or cost of promotion decreases. In any way, this means that enhancement over ROI is accomplished.

First of all, it is necessary to find income for each of 4 products and cost for each of 4 channels in the dataset. Profit and cost numbers are not provided as the dataset done. Therefore, research over literature is done to estimate these values. However, an approval on estimated values is received from a domain expert. Table 4.6 shows profit of each product and Table 4.7 demonstrates cost of each channel in the dataset. Note that profit and cost values are calculated in June, 2016 in Turkey. Monetary transaction amount is fixed to 1000 TRY (Turkish Liras).

4.2 Direct Marketing of Bank Products with Profit/Cost Analysis

In this section, there are 3 methods used over dataset with *IsSold* attribute. First method is based on C4.5 Decision Tree algorithm for classification to generate model over dataset. The model is used to decide whether customer accepts campaign offer or not. If result of model tree is *No*, then it means that no further time should be waste with *No* resulted customer. However, if the result

of model tree is *Yes*, then customer may accept the offer. The problem at this point is which product should be offered and which channel should be used for contact. To decide product and channel, clusters are used to find closest product and channel for that customer. Second method uses similar approach like first one; however, fast decision tree is used as a classification algorithm instead of decision tree. The last method uses same approach to decide whether customer accepts the offer or not, using classification model produced in first method. The most closest product is used for promotion as in first method. However, instead of using clustering to find channel, three different heuristic approaches are defined to find channel. After determining product with closest centroid, the most suitable channel or channels are assigned to the product that will be offered. First heuristic is based on assigning a channel to the best matched product. Second heuristic assigns two best matched channels for given product. Third one proposes the product with cheapest channels.

4.2.1 Profit/Cost Analysis over Model Based Bank Product and Channel Prediction Technique

This hybrid method is proposed in Section *Model Based Bank Product and Channel Prediction Technique* of papers [21] and [22]. This proposed method is improved applying profit/cost analysis. It consists of two phases, which are classification as the first phase and partitioning using clustering as the second phase. Therefore, it is called hybrid technique. In the first phase, classification model is used to decide whether customer will buy a product or not. This phase is a replacement of partitioning based method whose details can be found in paper [21] and [22]. In the second phase, closest centroid $P_iC_jY^{(k)}$ to the given customer C is found and P_i and C_j values are assigned as predicted values for customer C. Between methods that uses C4.5 decision tree and Naïve Bayes classifications, C4.5 decision tree used model based method has the best results. Therefore, C4.5 decision tree results are used to analyze profit/cost ratios over proposed method.

The main idea behind paper [21] and [22] is to increase the return rate of bank

marketing campaigns while increasing accuracy of prediction. In addition, specificity ratio which increases with true positive predictions, is as important as accuracy. However, a different approach is taken to this view and mentioned that profit/cost ratio is as important as accuracy and specificity rates. It is an important objective for banks to increase their profits of marketing campaign while decreasing cost of their campaigns.

In Algorithm 1, there are two phases applied. In phase 1, classification model is generated for given *dataset* and partitioning is applied on it. Each customer instance with different product and channel are distinguished and located in separated partitions. After that clustering is used on partitions, to increase possibility of that instances with common properties will remain in the same group. In phase 2, deciding whether customer accepts the offer or not, is evaluated using generated *model_tree*. If positive result is achieved after evaluation, closest cluster c in *cluster_list* to customer instance is found. After this stage, product of cluster c is proposed to customer via channel of cluster c. On the other hand, if negative result is achieved, no product is offered.

Profit/Cost Analysis over Model Based Bank Product and Channel Prediction Technique Using Fast Decision Tree

In this section, a modification on classification method of Section 4.2.1 is applied. In Algorithm 1, *model_tree* is generated using C4.5 Decision Tree. Instead of using C4.5 Decision Tree to produce *model_tree*, Fast Decision Tree Learner using REPTree implementation of WEKA is used. The motivation is that applying different methods on same data may change results of experiments.

On the other hand, REPTree can be used for regression analysis. However, because label attribute *IsSold* is nominal, regression analysis cannot be applied directly on dataset. It is required to change label attribute to numerical and this topic is mentioned in Section 4.3.

Algorithm 1 Model Based Bank Product and Channel Prediction Technique

Require: *dataset*: customer data result of campaign

```
procedure INITIALIZE_PHASE
//phase 1 - initialize prerequisite data
    partition_list ← {} // list of all partitions for dataset
    cluster_list ← {} // list of all clusters for partitions
    predict_list ← {} // list of all predictions for each customer
    train_data ← retrieve_train_data(dataset)
    test_data ← retrieve_test_data(dataset)
procedure CONSTRUCT_MODEL_AND_PARTITIONS
//phase 2 - construct model and partitions with clusters
    model_tree ← classify(train_data)
    for each instance i ∈ train_data do
        if i.Label is YES then
            add i to partition_list[i.Product][i.Channel]
    for each partition p ∈ partition_list do
        clusters ← apply_cluster_op(p, n) // n is cluster count
        add clusters to cluster_list
procedure MODEL_BASED_PREDICTION
//phase 3 - predict product, channel and label
    for each instance i ∈ test_data do
        if evaluation_result(model_tree, i) is YES then
            cluster ← find_closest_cluster(i, cluster_list) // distance function
            predict_list[i].Label ← YES
            predict_list[i].Product ← cluster.Product
            predict_list[i].Channel ← cluster.Channel
        else
            predict_list[i].Label ← NO
    return predict_list
```

Table 4.8: Channel Alignment

	Best Channel	Best Two Channels	Cheapest Channels
Credit Card (KK)	EMAIL	EMAIL, CC	EMAIL
Loan (KREDI)	CC	CC, SMS	SMS, EMAIL
Overdraft Account (KMH)	SMS	SMS, EMAIL	SMS, EMAIL
Deposit Account (VDL)	SMS	SMS, EMAIL	SMS, EMAIL

4.2.2 Profit/Cost Analysis over Model Based Bank Product and Channel Prediction Technique with Heuristic Support

To increase profit/cost ratio, a new method is proposed, which is an enhancement over method explained in Section 4.2.1. Rather than using distance calculation over all centroids of clusters to predict channel, three heuristics are added to predict channel, which are:

- Applying best channel for predicted product.
- Applying best two channels for predicted product.
- Applying cheapest channels for predicted product.

Table 4.9 explains details of aligning channels for each product. PCY combination holds only *Yes* class labeled customers and PCN one holds only *No* class labeled customers. Total number of PCY and PCN will be PCL which equals to whole dataset used for training. Totally, 73724 instances are used for training which is 90% of all dataset.

Clustering is used to find centroid of closest cluster to determine which product will be offered via which channel in previous methods. The modification in this section is that only product is found using clustering with this method. Channel is predicted with applying heuristics, which decrease time of process; however, there is no guarantee to increase accuracy. Nevertheless, well educated guesses may increase accuracy and profit/cost ratio. To make prediction process more faster and accurate, instead of using distance function to find channel, the channel/s which is/are most appropriate with predicted product is proposed.

Table 4.9: Success Rate for All Channels with PCL Combinations

PC Combination	Total PCY Count	Total PCN Count	PCY/ PCL (%)	Total Cost	Cost per Product
KK_IVN	14	1142	1.21	206.92	14.78
KMH_IVN	0	0	N/A	0	N/A
KREDI_IVN	46	8022	0.57	1444.17	31.39
VDL_IVN	0	0	N/A	0	N/A
KK_CC	793	7371	**9.71**	7347.6	9.26
KMH_CC	0	0	N/A	0	N/A
KREDI_CC	353	5462	**6.07**	5233.5	14.82
VDL_CC	0	0	N/A	0	N/A
KK_EMAIL	518	2905	**15.13**	3.42	0.01
KMH_EMAIL	296	7634	**3.73**	7.93	0.03
KREDI_EMAIL	92	8261	1.10	8.35	0.09
VDL_EMAIL	17	8788	**0.19**	8.80	0.51
KK_SMS	0	22	0.00	0.03	N/A
KMH_SMS	397	4888	**7.51**	7.39	0.02
KREDI_SMS	157	8136	**1.89**	11.61	0.07
VDL_SMS	19	8391	**0.23**	11.77	0.62

Table 4.8 shows the details of selected channels for each product used in heuristics.

Heuristic Support Applying Best Channel

In Algorithm 2, there are three phases; however, only third phase is demonstrated. Because first two phases of Algorithm 2 is same with first two phases in Algorithm 1. In phase 3, deciding whether customer accepts offer, is evaluated using generated *model_tree*. If result of model evaluation is positive, cluster c in *cluster_list* is found, where c is closest cluster to customer instance. Using cluster c, product field of c is proposed to customer. Main improvement of Algorithm 2 is applied at this step, to find which channel is used to offer the campaign of previously found product. Adding a function to find best channel is step by setting parameter *chan_num* to 1, and will detect best channel for previously predicted product. On the other hand, if negative result is achieved,

Algorithm 2 Model Based Bank Product and Channel Prediction Technique with Heuristic Support Applying Best Channels

Require: *chan_num*: number of best matched channel count

```
INITIALIZE_PHASE()                          ▷ phase 1 - initialize prerequisite data
CONSTRUCT_MODEL_AND_PARTITIONS()            ▷ phase 2 - construct model and partitions with clusters
procedure PREDICTION_WITH_HEURISTIC_SUPPORT
//phase 3 - predict product, channel and label
    for each instance i ∈ test_data do
        if evaluation_result(model_tree, i) is YES then
            cluster ← find_closest_cluster(i, cluster_list) // distance func
            predict_list[i].Label ← YES
            predict_list[i].Product ← cluster.Product
            best_channel ← find_best_channel(cluster.Product, chan_num)
            predict_list[i].Channel ← best_channel
        else
            predict_list[i].Label ← NO
    return predict_list
```

then no product is offered. Refer to Table 4.8 to see most successful channel.

Heuristic Support Applying Best Two Channels

In this section, there is a modification over previous section. Instead of proposing the channel which is most successful with predicted product, the most successful two channels are used to contact with customer.

Algorithm 2 is also applied in this section. Although, similar steps are applied as in previous heuristic, there is a minor change in this section. The difference is that added function to find best channel takes parameter *chan_num* as 2. Therefore, best two channels is used to offer previously predicted product using distance function. Refer to Table 4.8 to see most successful two channels.

Heuristic Support Applying Cheapest Two Channels

In this section, there is a modification over previous two sections with heuristics. Difference in this method is that the cheapest channels are used to offer predicted products.

Details of Algorithm 2 explains how the method in this section works. The only difference is to call function $find\_cheapest\_channel$ instead of function $find\_best\_channel$. Via the channel/s that are result of $find\_cheapest\_channel$, predicted product is offered to customer.

Refer to Table 4.8 to see the cheapest channel/s. Table 4.7 mentions about cost of each channel. Cheapest channels are selected using this information. Mention about Credit Card(KK) product, EMAIL channel is only use for cheapest channel because there does not exist any successful campaign result different than EMAIL for Credit Card in dataset.

4.3 Profit/Cost Analysis over Direct Marketing of Bank Products with Ratio Attribute

In this section, two different approaches are applied. First one uses regression tree to create model and clustering to group similar customer instances. Similar approach is applied as in Section 4.2.1. However, there is an important modification in this method. Instead of using decision tree, a new perspective is used for profit/cost analysis applying regression tree. Because regression tree needs numerical label, radical modifications on dataset is applied. Details are explained in Section 4.1.4. Second one is a combinational model generation. Three inner combination is applied over dataset to maximize profit/cost ratio while increasing accuracy as possible. Applied combinational models are:

- Model generation using product and channel combinations
- Model generation using channel combination
- Model generation using product combination

4.3.1 Profit/Cost Analysis over Direct Marketing of Bank Products using Regression Tree with Clustering

In this section, instead of predicting *Yes* or *No* values, a real number is generated using model tree and prediction is applied over this value. Instead of using information gain as in decision tree, means over sum of squared deviations are used to split the data.

Additional constraint should be supplied to predictor which is a line between positive and negative result. Therefore, minimum evaluation value called *threshold* is added inside of predictor to decide that prediction result is positive or not.

Algorithm 3 Regression Based Bank Product and Channel Prediction Technique

Require: *threshold*: minimum evaluation threshold for positive results

```
INITIALIZE_PHASE()                    ▷ phase 1 - initialize prerequisite data
procedure CONSTRUCT_MODEL_PARTITION
//phase 2 - construct model and partitions with clusters
    regression_tree ← classify(train_data)
    for each instance i ∈ train_data do
        if i.Label > 0 then
            add i to partition_list[i.Product][i.Channel]
    for each partition p ∈ partition_list do
        clusters ← apply_cluster_op(p, n) // n is cluster count
        add clusters to cluster_list
procedure REGRESSION_BASED_PREDICTION
//phase 3 - predict product, channel and label
    for each instance i ∈ test_data do
        if evaluation_result(regression_tree, i) > threshold then
            cluster ← find_closest_cluster(i, cluster_list) // distance func
            predict_list[i] ← {YES, cluster.Product, cluster.Channel}
        else
            predict_list[i].Label ← {NO, NULL, NULL}
    return predict_list
```

In Algorithm 3, same initialization procedure as in Algorithms 1 and 2, is called to initialize dataset and required data structures.

In phase 2, a regression model is generated using training data. Then, training data is spitted into partitions where each partition represents PC combination. After that, each partition is divided into subgroups by applying clustering operations. The important difference in this phase is that generated model is regression tree and decision of whether an instance will be in a partition or not, is made using numerical condition check. If label value of the instance is more than 0, then instance is assumed to be valuable. If it is 0 or less, then instance is checked as negative and will not placed into a partition.

In phase 3, *threshold* concept is presented, where evaluation result of regression model result is checked with it. If evaluation result is more than *thresold*, then input instance of regression model is predicted as *Yes*. Product and channel assignment is done using distance function on clusters, where closest cluster to instance is selected to use its product and channel. If evaluation result is less than *thresold*, then the instance is assumed as negative and labeled as *No*.

4.3.2 Profit/Cost Analysis over Direct Marketing of Bank Products with Ratio Attribute with Combinational Model Generation

In this section, a new technique is proposed which is a modification of Section 4.3.1. Normally, a regression model is generated on training data; however, model generation is applied on partitions and more than one model is created. There are three methods that are used for combinational model generation, which are as below:

- Combination of product and channel model generation
- Combination of channel model generation
- Combination of product model generation

Combination of Product and Channel Model Generation

This model generation method produces partitions with total size of combination

of products and channels. In dataset, there exist 4 products and 4 channels; therefore, 16 partitions are produced. For each partition, one regression model is generated using regression tree classifier.

Algorithm 4 Regression Based Bank Product and Channel Prediction Technique with Product and Channel Model Generation

Require: *threshold*: minimum evaluation threshold for positive results

```
INITIALIZE_PHASE()                    ▷ phase 1 - initialize prerequisite data
procedure CONSTRUCT_MODEL_PARTITION
//phase 2 - construct model and partitions
    for each instance i ∈ train_data do
        add i to partition_list[i.Product][i.Channel]
    product_list ← get_product_list(train_data)
    channel_list ← get_channel_list(train_data)
    size ← sizeof(product_list)
    for each product p ∈ product_list do
        for each channel c ∈ channel_list do
            regression_tree[size * p + c] ← classify(partition_list[p][c])
procedure COMBINATIONAL_MODEL_BASED_PREDICTION
//phase 3 - predict product, channel and label
    for each instance i ∈ test_data do
        pre_eval.Value ← 0 // result of model tree for an instance
        for each partition p ∈ partition_list do
            current_eval ←evaluation_result(regression_tree[p], i)
            if current_eval.Value > pre_eval.Value then
                pre_eval.Value ← current_eval.Value
                pre_eval.Product ← current_eval.Product
                pre_eval.Channel ← current_eval.Channel
        if pre_eval.Value > threshold then
            predict_list[i] ← {YES, pre_eval.Product, pre_eval.Channel}
        else
            predict_list[i].Label ← {NO, NULL, NULL}
    return predict_list
```

In Algorithm 4, details of model generation with combination of product and channel, are explained.

In phase 3, each instance of test data is evaluated with 16 models and the highest result with its model is selected after evaluation step. The selected evaluation value is checked with *threshold* and if it is higher than *threshold*, then product and channel of partition selected with evaluation is assigned to instance with positive prediction. If evaluation value is less than *threshold*, then instance is marked as negative and not included into prediction list.

Combination of Channel Model Generation

In this section, models are generated using channel list, which contains all channels used in dataset. Similar approach is followed as in previous method. However, only channels are used to create partitions and for each of them, separate partitions are generated.

In this training data, there are 4 channels. Therefore, 4 partitions are generated and each partition is classified using regression tree to build models. In phase 3, for each instance $i \in test\_data$, i is evaluated using all regression models and the evaluation result *eval* with highest one is selected as prediction candidate. If *eval* is greater than *threshold*, then i is added to prediction list as positive instance. Channel of contact is selected from partition which is used to generate successful model. The handicap of this method is that positively predicted customer is contacted for all products in *product_list*. There are two ways to solve this problem, which are:

- Proposing all products in the same call
- Proposing all products in separate calls

If *eval* value is less than *threshold* means that customer is not worth to be contacted. Therefore, it is added to *predict_list* as negative instance.

Combination of Product Model Generation

In this section, models are generated using product list, which contains all prod-

ucts used in dataset. Same approach is applied as in previous section. Instead of generating partitions using channels, products are used to create partitions. Therefore, one regression model for each partition is generated.

In the dataset, there are 4 products. Therefore, 4 partitions are generated and each partition is classified using regression tree. If *eval* which is highest result over all models is greater than *threshold*, then i is added to prediction list as positive instance. Product of partition constituting the selected model, is assigned to offer. If *threshold* is greater than *eval* value, then customer is not valuable to be contacted for a product. Therefore, it is added to prediction list with negative value.

The disadvantage of this method is that there is no any judgment for which channel is used to offer the selected product. Therefore, all channels are used to proposed the product.

CHAPTER 5

RESULTS AND DISCUSSIONS

In this chapter, results of profit/cost analysis over direct marketing of bank products are discussed. Content is separated into sections as used in 4. First section is results of experiments with methods on *IsSold* attribute. Results of hybrid method and fast decision tree modification is detailed in this part. Then, experiments on dataset with *Ratio* attribute are detailed with second section. Results of regression analysis and combinational regression model are detailed at this stage.

In experiments, WEKA Data Mining Software [13] v3.6 is used to train and test the dataset. Methods proposed in this thesis are implemented using Java and Java API of WEKA, and they are run using personal computer with Intel® Core™ i5-5200U with 2 physical 2.20 GHz CPUs and physical memory with capacity 8 GB.

5.1 Baseline Method for Profit/Cost Analysis

To compare test results, there should be a starting point for comparison. Therefore, an appropriate baseline is generated for the dataset focused on profit and cost constraints. Each customer is assumed to be contacted for the product via the channel as specified in the dataset.

Results shows that real cost of our test data is 1654.25 TRY and real profit of our test data is 8785 TRY. This means that it would have been achieved maximum 8785 TRY profit using this test dataset if all customers with *Yes* class has been

predicted correctly. On the other side, if all customers had been contacted to promote the product via specified channel to get maximum profit, the maximum cost value would have been achieved. Therefore, using Equation 4.2, profit/cost ratio of real dataset is found as 5.31.

5.2 Direct Marketing of Bank Products with Profit/Cost Analysis

This section contains experiments of two different techniques, which are hybrid and heuristic approach, respectively. Hybrid approach uses two different classification algorithms and three alternative heuristics are defined in heuristic approach.

5.2.1 Profit/Cost Analysis over Model Based Bank Product and Channel Prediction Technique

In this experiment, two different classification methods are used, which are C4.5 decision tree and fast decision tree learners. Compared to baseline model, profit/cost ratio increases with both of methods. Table 5.1 shows that prediction technique with C4.5 decision tree has 7.70 profit/cost ratio with 3 clusters. On the other hand, after modifying the prediction technique with fast decision tree algorithm, profit/cost ratio increases up to 7.77 with 3 clusters. It would have been expected to get maximum profit/cost ratio with 100 clusters; however, there is an optimal cluster count for a dataset to distribute its data to clusters and in this experiments, the best results that are obtained for profit/cost ratio is 3 clusters. On the other hand, accuracy and true positive rates increase proportional to cluster count and the maximum values are obtained with 100 clusters. Therefore, it can be said that profit/cost ratio and accuracy rate are not change proportionally. Because, experiments show that profit/cost ratio increases proportional to cluster count until reaching specific cluster count for each dataset. Additionally, fast decision tree learner has better results than C4.5 decision tree learner, referred to Table 5.1.

Table 5.1: Analysis with C4.5 and Fast Decision Tree

num. of clusters	Acc. (%)	TP Count	TP (%)	Profit	Cost	Profit-Cost Rate	
1	72.61	31	28.44	1083	**253.4**	4.27	C4.5 decision tree
3	73.76	125	61.57	3602	467.4	**7.70**	
5	73.87	134	63.20	3925	547.2	7.17	
10	73.82	130	62.50	4022.5	564.8	7.12	
20	74.09	152	66.08	4739	707.4	6.69	
40	74.19	160	67.22	4947.5	713.4	6.93	
100	**74.47**	**183**	**70.11**	**5579**	742	7.51	
1	74.11	31	32.29	1077.5	**225.9**	4.76	Fast decision tree
3	75.38	135	67.5	3967.5	510	**7.77**	
5	75.49	144	68.89	4278	624.8	6.84	
10	75.42	138	67.98	4283.5	608	7.04	
20	75.76	166	71.86	5224.5	765.9	6.82	
40	75.83	172	72.57	5396.5	763.3	7.06	
100	**76.14**	**197**	**75.19**	**6088**	785.3	7.75	

5.2.2 Profit/Cost Analysis over Model Based Bank Product and Channel Prediction Technique with Heuristic Support

In this experiment, 3 heuristic methods, which are *Best Channel*, *Best Two Channels* and *Cheapest Channel/s* are implemented and results are in Table 5.2. All three of heuristics have advantages and disadvantages. There can not be made exact judgment that one of them is best. However, results show that *Cheapest Channel* has highest profit/cost ratio with 136.10. Because, the cheapest channels used, cost of campaigns decreases dramatically. On the other hand, the best accuracy rate with highest correctly predicted positive customer count is achieved 74.47 with *Best Two Channels* using 100 clusters. In 100 clusters, this result is same with experiment of method using C4.5 decision tree learner in Table 5.1. However, cost and profit/cost ratio have different values. Because heuristics may use more than one channel to contact with customer, cost value increases and causes profit/cost rate go down.

Table 5.2: Analysis with 3 Heuristics Based on Channels

num. of clusters	Acc. (%)	TP Count	TP (%)	Profit	Cost	Profit-Cost Rate	
1	72.62	32	29.09	1233.5	752.3	1.63	Best channel
3	73.42	97	55.42	3042	805.8	3.77	
5	73.42	97	55.42	3042	799.3	3.80	
10	73.29	87	52.72	2876	813.3	3.53	
20	73.31	88	53.01	2911.5	749.9	3.88	
40	73.29	87	52.72	2863.5	755.1	3.79	
100	**73.53**	**106**	**57.60**	**3397**	**747.8**	**4.54**	
1	72.76	43	35.53	1564	**1042.1**	1.50	Best two channels
3	73.88	135	63.38	4009	1563.9	2.56	
5	73.97	142	64.54	4254	1516	2.80	
10	73.88	135	63.38	4206	1356.7	3.10	
20	74.10	153	66.23	4763.5	1269.8	3.75	
40	74.20	161	67.36	4972	1284.9	3.86	
100	**74.47**	**183**	**70.11**	**5579**	1352.2	**4.12**	
1	72.59	29	27.10	1095	28.62	38.24	Cheapest channels
3	73.21	80	50.63	2319.5	**21.34**	108.65	
5	73.27	85	52.14	2515.5	21.99	114.38	
10	73.16	76	49.35	2364.5	24.69	95.75	
20	73.25	83	51.55	2616.5	25.04	104.47	
40	73.28	86	52.43	2738.5	24.89	110.02	
100	**73.50**	**104**	**57.14**	**3229.5**	23.72	**136.10**	

5.3 Profit/Cost Analysis over Direct Marketing of Bank Products with Ratio Attribute

This section contains experiments over two different methods; hybrid approach using regression tree as a classifier and combinational model generations. In method with regression tree, two different experiments are applied which depend on changing cluster count and threshold while one of them is fixed. In combinational model generation method, three different approach is followed.

5.3.1 Profit/Cost Analysis over Direct Marketing of Bank Products using Regression Tree with Clustering

In this experiment, results of regression tree learner implementation on dataset with ratio attribute is demonstrated. Table 5.3 is based on changing cluster count while threshold value is fixed to 100. Table 5.4 is generated by changing threshold value while cluster count is fixed to 100.

Table 5.3: Regression Analysis with Ratio Attribute Using Cluster Count

num. of clusters	Acc. (%)	TP Count	TP (%)	Profit	Cost	Profit-Cost Rate	
1	38.85	38	23.03	1371	719.5	1.90	Change cluster count with threshold 100
3	39.53	93	42.27	2790.5	314.2	8.88	
5	39.60	99	43.80	3070.5	317.3	9.67	
10	39.54	94	42.53	3084	287.5	10.72	
20	39.70	107	45.72	3569	291	12.26	
40	39.83	118	48.16	3960.5	299	13.24	
100	**40.03**	**134**	**51.34**	**4408**	**231.5**	**19.04**	

Table 5.4: Regression Analysis with Ratio Attribute Using Threshold

Threshold	Acc. (%)	TP Count	TP (%)	Profit	Cost	Profit-Cost Rate	
50	21.20	**188**	**74.30**	**5785**	1193.2	4.84	Change threshold with 100 clusters
100	40.03	134	51.34	4408	231.5	19.04	
250	41.03	133	50.95	4365	230	18.97	
400	44.48	129	49.42	4217.5	225.4	18.71	
1000	67.98	144	42.69	3621.5	184.2	**19.66**	
2000	**72.56**	107	39.92	3362.5	**180.6**	18.61	

Table 5.3 shows that increasing cluster count has not significant effect on accuracy; however, because true positive prediction rate increases, propositionally profit/cost ratio increases too. With fixed threshold, larger cluster count positively affects profit/cost ratio. However, false positive count is high. This can be

solved by increasing threshold value and true negative value can be enhanced. Table 5.4 shows that increasing threshold with constant cluster count enhances accuracy. Additionally, increasing threshold decreases cost because false positive value decreases.

5.3.2 Profit/Cost Analysis over Direct Marketing of Bank Products with Ratio Attribute with Combinational Model Generation

In this experiment, experiments on regression analysis with combinational model generation is demonstrated. Table 5.5 has 4 parts, which are *product combination*, *channel combination* with 2 different implementations and *product and channel combinations*. About *channel combination*, first implementation promotes all products in same call and second one promotes products in separate calls.

Results show that threshold should be adjusted carefully and optimal value for it should be found. Profit/cost ratio by itself is not confidential to work on. Although, it increases while accuracy increases, only true negative count goes up. True positive values yield to decrease and converges to 0. Because, negative response count is approximately 27 times greater than positive ones, losing true positive does not affect accuracy too much.

Table 5.5: Regression Analysis with Ratio Attribute Using Combinational Model Generation

Threshold	Acc. (%)	TP Count	TN Count	FP Count	FN Count	Profit	Cost	Profit-Cost Rate	
0	0.72	59	0	8132	0	1620.5	8.23	196.90	combined 16 trees
1	0.72	59	0	8132	0	1620.5	8.23	196.90	
10	0.72	59	0	8132	0	1620.5	8.23	196.90	
100	0.72	59	0	8132	0	1620.5	8.23	196.90	
1000	14.32	55	1118	6999	19	1472.5	7.09	207.62	
10000	36.01	42	2908	5170	71	1079	5.21	207.10	
30000	96.69	3	7917	1	270	111	0.01	27750	
0	1.26	104	0	8087	0	3330.5	8960.9	0.37	4 product trees
1	1.26	104	0	8087	0	3330.5	8960.9	0.37	
10	1.40	104	11	8076	0	3330.5	8948.9	0.37	
100	12.2	97	906	7167	21	3066	7946.8	0.38	
1000	61.39	59	4970	2996	116	1588.5	3342.17	0.47	
10000	89.15	31	7272	662	226	809.5	758.1	1.06	
30000	96.67	3	7916	3	269	111	6.56	16.92	
0	1.40	114	1	8076	0	3517	578.4	6.08	4 channel trees
1	1.46	114	6	8071	0	3517	578.4	6.08	
10	16.10	99	1220	6856	16	3077.5	24.3	126.64	
100	37.87	93	3009	5026	63	2822.5	5.56	507.64	
1000	83.46	76	6761	1188	166	2155	1.5	1436.66	
10000	96.04	28	7839	80	244	736	0.108	6814.81	
30000	96.66	0	7918	0	273	0	0	0	
0	1.40	114	1	8076	0	3517	2313.3	1.52	4 channel trees with separate calls
1	1.46	114	6	8071	0	3517	2313.3	1.52	
10	16.10	99	1220	6856	16	3077.5	97.3	31.62	
100	37.87	93	3009	5026	63	2822.5	22.24	126.91	
1000	83.46	76	6761	1188	166	2155	6	359.16	
10000	96.04	28	7839	80	244	736	0.43	1711.62	
30000	96.66	0	7918	0	273	0	0	0	

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this section, methods that are applied over data mining on direct marketing with profit/cost ratio are concluded. Then, future works are explained which may be implemented over works of this thesis.

6.1 Conclusion

This thesis contains two different datasets, which consist of *IsSold* and *Ratio* labels separately. With respect to datasets, different methods are applied for profit/cost analysis.

For dataset with *IsSold* class label, a proposed method in [21] is implemented. Firstly, classification is applied on dataset to generate model tree to determine whether customer yields to accept the promotion or not. Secondly, clustering is used on dataset to group instances according to similar characteristics to determine communication channel and product to promote. Then, a modification is applied on classification method of this hybrid method and fast decision tree is used instead. Additionally, new heuristics are defined to determine communication channels. As before, model tree is generated on dataset using classification and clustering the data instances is applied to find the product to promote. Newly, channels are found using predefined heuristics, which are *Best channel*, *Best two channels* and *Cheapest channel/s*.

Compared to profit/cost ratio of baseline method which has 5.32, 45% of increase is achieved on hybrid method with C4.5 Decision tree classifier. On the other

hand, after applying fast decision tree, ratio increases from 7.70 to 7.77 compared to previous classifier. Profit/cost ratio of fast decision tree is 46% greater than baseline method ones. Although, there is a few change over ratio between two classifier implementation, accuracy increases from 74.47% to 76.14%, which provides 9.12% more profit in 100 clusters in second classifier.

Between 3 heuristic approaches, the best profit/cost ratio is achieved with *Cheapest channel/s* heuristic, which is 136.10. In terms of profit/cost ratio, *"cheapest channel/s > best channel > best two channels"* rule is acquired. Although, total profit of *Cheapest channel/s* heuristic (3229.5) is lower than *Best two channels* one (5579), it should be mentioned that high ratio in *Cheapest channel/s* heuristic is achieved because of low cost of channels which are EMAIL and SMS. In addition, in 100 clusters, profit value of best two channels is same with hybrid approach with C4.5 decision tree classifier. However, because of using heuristic instead of clusters for channel prediction, cost value of heuristic one is twice as much as cluster used one.

Results shows that the highest profit/cost ratio is achieved with *Cheapest channel/s* heuristic with 136.10. However, because of the greatest true positive rate is succeeded in hybrid method with fast decision tree, the highest profit is achieved with value of 6088 too.

Original dataset is upgraded using profit and cost estimations. Using them, *Ratio* attribute is replaced with *IsSold* one. On new dataset, two different approaches are applied. Firstly, regression tree is used as a classifier and *threshold* is added to decide offering promotion. Any instance with *Ratio* value is greater than 0 is assumed to be positive and placed in one of clusters. In testing phase, if model result is greater than threshold, then the cluster with closest distance to data instance is used to decide which product will be promoted via which channel to customer. Secondly, a different approach is applied, which is combinational model generation. 3 separate models are generated, which are *product combination*, *channel combination* and *product and channel combinations*. Dataset is partitioned according to combination type and then models with total number of combination count is generated. A data instance is evaluated for each gen-

erated model and if the highest one is greater than threshold, then the product and channel of partition belongs to model is used for promotion.

Results of regression tree show that increasing threshold decreases true positive rate. Therefore, it should be supported by increasing cluster count. The best profit/cost ratio (19.04), is achieved while threshold is fixed to 100 and cluster count is maximized to 100. Compared to baseline method result, new method has 358% greater profit/cost ratio. After this step, cluster count is fixed to 100 clusters and effects of increasing threshold is experimented. 19.66 profit/cost ratio is achieved with threshold 1000 and this result is 370% greater than baseline method one. Although increasing threshold enhances profit/cost ratio and accuracy, it is important to mention that true positive ratio and total profit decrease while threshold increases. Therefore, optimal cluster count and threshold should be found for each dataset.

Profit/cost ratio of combinational model generations are greater than all other methods; however, results may misdirect someone. Although high ratio is achieved, accuracy and true positive count are generally too low. The optimal results are achieved with channel combination model using threshold 1000, where profit/cost ratio is 1436.6 with accuracy 83.46%. However, only 76 of all positive instances are predicted correctly which is 27.83% of all positive data instances in test dataset. High accuracy is achieved because of true negative predictions. Since it is used only classification in this methods and results are not as good as previous methods, it is important to say that applying hybrid approach will give better results.

Results shows that the highest profit/cost ratios are achieved with combinational model generation. However, they are not feasible to apply on banking campaigns. On the other hand, more feasible results are achieved with regression tree using threshold 1000. With 19.66% profit/cost ratio, the best results are succeeded using regression tree over dataset with *Ratio* attribute.

The fact is that one should decide which constraint is important for the campaign. Because there is a trade-off between accuracy, profit/cost ratio, total profit value, total cost value, true positive and true negative values, it is not

possible to get the best result for all constraint with single prediction.

6.2 Future Work

Accuracy rate and profit/cost ratio depend on datasets. Methods that are proposed in this thesis can be applied on different datasets and results can be compared. Additionally, a dataset with 13 attributes is used in this thesis. Increasing attribute count, which discriminates data instances more precisely, may enhance accuracy, and improve profit/cost ratio.

Additionally, proposed methods can be applied not only bank datasets, but also different dataset belongs to finance, supermarket, insurance, health-care, automotive, etc.

Moreover, classification methods used in combination model generation can be supported with clustering to improve results. Additionally, results of proposed methods can be enhanced by changing their clustering and classification methods. It is proved with hybrid methods that each method gives different results with same dataset.

Finally, this work can be improved using ensemble methods like bagging and boosting algorithms.

REFERENCES

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] T. F. Bahari and M. S. Elayidom. An efficient crm-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, 46:725–731, 2015.

[3] C. L. Bauer and J. Miglautsch. A conceptual definition of direct marketing. *Journal of Direct Marketing*, 6(2):7–17, 1992.

[4] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[5] W.-y. K. Chiang, D. Chhajed, and J. D. Hess. Direct marketing, indirect profits: A strategic analysis of dual-channel supply-chain design. *Management science*, 49(1):1–20, 2003.

[6] M.-D. Cohen. Exploiting response models—optimizing cross-sell and up-sell opportunities in banking. *Information Systems*, 29(4):327–341, 2004.

[7] H. A. Elsalamony. Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications*, 85(7), 2014.

[8] R. Elsner, M. Krafft, and A. Huchzermeier. Optimizing rhenania's direct marketing business through dynamic multilevel modeling (dmlm) in a multicatalog-brand environment. *Marketing Science*, pages 192–206, 2004.

[9] P. Engardio and D. Roberts. The china price". *Business Week*, pages 42–49, 2004.

[10] E. Frank and B. Pfahringer. Improving on bagging with input smearing. In *PAKDD*, volume 3918, pages 97–106. Springer, 2006.

[11] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[12] C. Giraud-Carrier and O. Povel. Characterising data mining software. *Intelligent Data Analysis*, 7(3):181–192, 2003.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[14] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[15] W. Kamakura, C. F. Mela, A. Ansari, A. Bodapati, P. Fader, R. Iyengar, P. Naik, S. Neslin, B. Sun, P. C. Verhoef, et al. Choice models and customer relationship management. *Marketing Letters*, 16(3):279–291, 2005.

[16] P. Kang, S. Cho, and D. L. MacLachlan. Improved response modeling based on clustering, under-sampling, and ensemble. *Expert Systems with Applications*, 39(8):6738–6753, 2012.

[17] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

[18] P. Kotler. From mass marketing to mass customization. *Planning review*, 17(5):10–47, 1989.

[19] C. X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *KDD*, volume 98, pages 73–79, 1998.

[20] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[21] M. Mitik, O. Korkmaz, P. Karagoz, I. H. Toroslu, and F. Yucel. Data mining based product marketing technique for banking products. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 552–559. IEEE, 2016.

[22] M. Mitik, O. Korkmaz, P. Karagoz, I. H. Toroslu, and F. Yucel. Data mining approach for direct marketing of banking products with profit/cost analysis. *The Review of Socionetwork Strategies*, pages 1–15, 2017.

[23] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[24] J. Murrow and M. R. Hyman. Direct marketing: Passages, definitions, and déja vu. *Journal of Direct Marketing*, 8(3):46–56, 1994.

[25] A. Nachev and M. Hogan. Application of multilayer perceptrons for response modeling. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014.

[26] E. W. Ngai, L. Xiu, and D. C. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[27] C. Ou, C. Liu, J. Huang, and N. Zhong. On data mining for direct marketing. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pages 491–498. Springer, 2003.

[28] Y. Pan and Z. Tang. Ensemble methods in bank direct marketing. In *Service Systems and Service Management (ICSSSM), 2014 11th International Conference on*, pages 1–5. IEEE, 2014.

[29] T. Piton, J. Blanchard, and F. Guillet. Capre: A new methodology for product recommendation based on customer actionability and profitability. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 466–473. IEEE, 2011.

[30] A. Prinzie and D. Van den Poel. Constrained optimization of data-mining problems to improve model performance: A direct-marketing application. *Expert Systems with Applications*, 29(3):630–640, 2005.

[31] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[32] J. R. Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[33] J. R. Quinlan. C4.5: Programming for machine learning. *Morgan Kauffmann*, 38, 1993.

[34] J. R. Quinlan. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4:77–90, 1996.

[35] R. Quinlan. Data Mining Tools See5 and C5.0, 2004.

[36] M. Raaijmaakers, J. Hoekstra, P. Leeflang, and M. Wedel. Success of communication strategies. In *Marketing for Europe—Marketing for the Future, Proceedings of the 21st Annual Conference of the European Marketing Academy, KG Grunert and D. Fuglede eds., Aarhus, Denmark*, pages 1383–1386, 1992.

[37] L. Rokach and O. Maimon. Decision trees. *Data mining and knowledge discovery handbook*, pages 165–192, 2005.

[38] B. Tedeschi. Compressed data; big companies go slowly in devising net strategy. *New York Times*, 26(1), 2000.

[39] A. R. Thomas. The end of mass marketing: or, why all successful marketing is now direct marketing. *Direct Marketing: An International Journal*, 1(1):6–16, 2007.

[40] H. van de Scheer. *Quantitative approaches for profit maximization in direct marketing*. Rijksuniversiteit Groningen, 1998.

[41] R. S. Winer. A framework for customer relationship management. *California management review*, 43(4):89–105, 2001.

[42] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[43] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.